IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Before the Board of Patent Appeals and Interferences

In re the Application

Inventor	:	Mingkun Li et al.
Application No.	:	10/076,194
Filed	:	February 14, 2002
For	:	METHOD AND SYSTEM FOR PERSON IDENTIFICATION USING VIDEO-SPEECH MATCHING

APPEAL BRIEF

On Appeal from Group Art Unit 2626

Daniel Piotrowski
Registration No. 42,079

Date: April 16, 2007

By: Steve Cha
Attorney for Applicant
Registration No. 44,069

TABLE OF CONTENTS

TABLE OF CASES

I.   REAL PARTY IN INTEREST

The real party in interest is the assignee of the present application, U.S. Philips Corporation, and not the party named in the above caption.

II.   RELATED APPEALS AND INTERFERENCES

With regard to identifying by number and filing date all other appeals or interferences known to Appellant which will directly effect or be directly affected by or have a bearing on the Board's decision in this appeal, Appellant is not aware of any such appeals or interferences.

III.   STATUS OF CLAIMS

Claims 1-2 and 4-20have been presented for examination. All of these claims are pending, stand finally rejected, and form the subject matter of the present appeal.

IV.   STATUS OF AMENDMENTS

In response to the Final Office Action, having a mailing date of November 15, 2006, applicant timely submitted arguments to overcome the reasons for rejecting the claims. Amendments were made to the claims 1, 8 and 15. In reply, an Advisory Action, having a filing date of February 15, 2007, was entered into the record. The Advisory Action stated that for purposes of appeal the amendments made to the claims would be entered. A copy of the claims, in final form, is shown in the Claims Appendix, below.

The Advisory Action in addition provided further rationale for maintaining the rejection of the claims

A Notice of Appeal was timely filed in response to the Advisory Action and this Appeal Brief is being filed, with appropriate fee, within the period of response from the date of the Notice of Appeal

V.   SUMMARY OF CLAIMED SUBJECT MATTER

The present invention is expressed primarily in independent claims 1, 8 and 15.

Independent claim 1 recites an audio-visual system for identifying a speaking person from video data comprising an object detection model (item 30, Figure 1) to provide object features from the video data , the object features selected from the group of temporal and spatial features, an audio processor (item 20, Figure 1) for providing audio features from the video data and a processor (item 100, Figure 2) for determining a maximum correlation value (see page 13, Equation 7) among a plurality of correlation values wherein each of the correlation values is determined as the sum of the correlation of selected object features and audio features (see page 13, lines 9-14).

Claims 1 and 15 recite a method (see page 19, lines 9-22) and a memory medium (see page 16, lines 25-26) for processing a video signal for identifying a speaking person by determining a maximum correlation value among a plurality of correlation wherein each of the correlation values is determined as the sum of the correlation of selected object features and audio features.

The remaining claims, which depend from respective independent claims, express further aspects of the invention.

VI. GROUNDS FOR REJECTION TO BE REVIEWED ON APPEAL

The issues in the present matter are whether:

1. Claims 1, 2, 4, 5, 8, 11 and 16-17 are unpatentable over Basu (USP no. 6,219,640) in view of Nevenka (USPPA 2003/0108334) under §35 USC 103(a); and

2. Claims 6 and 7 are unpatentable over Basu in view of Nevenka and further in view of Bradford (USPPA 2002/0103799) under §35 USC 103(a).

VII. ARGUMENT

I. 35 USC §102 Rejection of claims 1, 2, 4, 5, 8, 11 and 16-17

The rejection of claims 1, 2, 4, 5, 8, 11 and 16-17 as being rendered obvious under 35 USC §103(a) by the combination of Basu in view of Nevenka is in error because the references, when combined, fail to show a limitation cited in the independent claims.

**Difference between the Claimed Invention
as Recited in the Independent Claims
and the Cited References**

The instant invention, as recited in claim 1, for example, which is typical of the remaining independent claims, discloses a system for determining a speaker in a video image by determining a maximum correlation value among a plurality of correlation values, wherein each of the correlation values is determined based on correlation of subsets of audio and video information.

Basu describes a method and apparatus for performing speaker recognition comprising processing a video signal and an audio signal using a score combination

approach, a feature combination approach and a re-scoring approach. In one aspect Basu teaches that video and audio signals are separately processed and "the top three scores from the face identification process may be combined with the top three scores from the acoustic speaker identification process. Then the highest combined score is identified to the speaker." (see col. 10, lines 7-11). In this aspect, Basu teaches processing the video features separately from the audio features and then processing the top three scores of each of these identified video and audio aspects.

In another aspect, (see col. 12, lines 21-22), Basu teaches that audio and video features may be combined into an AV feature vector using a linear interpolation from frames immediately preceding and following the time instant (see col. 12, lines 33-48). A decision is made based on the AV feature vector having the highest score.

Nevenka is proposed for teaching, in part, that audio elements may be composed of low level elements of bandwidth, energy and pitch. The Advisory Action, states that "Nevenka do[es] teach audio parameters which would have correlation value in order to match the audio parameters such as energy ... page 6, paragraph 65 lines 9-11."

**Proposed Modification of Basu by Nevenka
Fails to Arrive at the Present Invention
as Recited in Claim 1**

In order to establish a *prima facie* case of obviousness, three basic criteria must be met;

> 1. there must be some suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to modify the reference or combine the reference teachings;
>
> 2. there must be a reasonable expectation of success; and

3. the prior art reference must teach or suggest all the claim limitations. The teaching or suggestion to make the claimed combination and the reasonable expectation of success must be found in the prior art, and not based on applicant's disclosure. *In re Vaeck*, 947 F.2d 488, 20 USPQ2d 1438 (Fed. Cir. 1991).

The rejection of the claims fails to address the claim elements "correlation values are determined from selected audio features and each of the video features" or "that a sum of the correlation values associated with each video feature is used to determine a maximum correlation value."

Basu fails to teach determining correlation values associated with audio and visual elements and a determination of a maximum correlation of the determined correlation values.

Nevenka fails to teach or suggest correlation values are determined from selected audio features and video features or that a sum of the correlation values associated with is used to determine a maximum correlation value.

With regard to the subject matter recited in claim 1, Applicant respectfully submits that pursuant to the three basic criteria a *prima facie* case of obviousness has not been set forth because the combination of the cited references fails to address a material element claimed.

The Manual of Patent Examining Procedure (MPEP) provides further appropriate instruction by which the instant Appeal should be judged. MPEP, Eight Edition, Rev. 2, May 2004, provides in section 2143 entitled: "Fact That The Claimed Invention Is Within The Capabilities Of One Of Ordinary Skill In The Art Is Not Sufficient By Itself To Establish *PRIMA FACIE* Obviousness:"

"A statement that modification of the prior art to meet the
claimed invention would have been "'well within the
ordinary skill of the art at the time the claimed invention
was made'" because the references relied upon teach that
all aspects of the claimed invention were individually
known in the art is not sufficient to establish a prima facie
case of obviousness without some objective reason to
combine the teachings of the references." *Ex parte
Levengood* 28 USPQ2d 1300 (Bd. Pat. App. & Inter. 1993).
MPEP §2143.01, p. 2100-131.

Appellant respectfully submits that the Office has failed to show the motivation

for modifying the teaching of Basu to perform correlation using elements of the audio

and visual objects.

For at least the above reasons, Appellant respectfully submits that a case of

obviousness has not been set forth.

In view of the above, applicant submits that claim 1 is not obvious in view of the

teachings of the cited references.

With regard to the remaining independent claims, these claims recite subject

matter similar to that recited in claim 1 and have been rejected citing the same references

as those recited in rejecting claim 1. Hence, the arguments presented in response to the

rejection of claim 1, herein, are applicable to the rejection of the remaining claims and

reasserted, as if in full, herein.

For the arguments presented herein, applicant submits that the combination of

Basu and Nevenka cannot render obvious the remaining independent claims, as the

combination of Basu and Nevenka fails to disclose every element recited in the remaining

independent claims.

With regard to the remaining claims, these claims depend from the independent claims. Applicant respectfully submits that these claims are allowable at least for their dependence upon allowable base claims, without even contemplating the merits of the dependent claims for reasons analogous to those held in *In re Fine*, 837 F.2d 1071, 5 USPQ 2d 1596 (Fed. Cir. 1988) (if an independent claim is non-obvious under 35 U.S.C. §103(a), then any claim depending therefrom is non-obvious).

In view of the above, applicant submits that the independent claims and the claims dependent therefrom are not rendered obvious over the teaching of the cited references.

2. 35 USC §103 Rejection of claims 6 and 7

The rejection of claims 6 and 7 as being rendered obvious under 35 USC §103(a) by the combination of Basu, Nevenka and Bradford is in error because the references, when combined, fail to show a limitation cited in independent claim 1 from which claims 6 and 7 depend.

Claims 6 and 7 Depend From an Allowable Base Claim

Claims 6 and 7 depend from independent claim 1, which has been shown to include subject matter not disclosed by the combination of Basu and Nevenka. Bradford fails to provide any teaching to correct the deficiency found to exist in Basu and Nevenka.

Applicant respectfully submits that claims 6 and 7 are allowable at least for their dependence upon an allowable base claim for the reasons held in *In re Fine*, 837 F.2d 1071, 5 USPQ 2d 1596 (Fed. Cir. 1988) (if an independent claim is non-obvious under 35 U.S.C. §103(a), then any claim depending therefrom is non-obvious).

9

In view of the above, applicant submits that the above referred-to claims are patentable over the teachings of Basu, Nevenka and Bradford.

VIII. CONCLUSION

In view of the above analysis, it is respectfully submitted that the referenced teachings, whether taken individually or in combination, fail to anticipate or render obvious the subject matter of any of the present claims. Therefore, reversal of all outstanding grounds of rejection is respectfully solicited.

Respectfully submitted,
Daniel Piotrowski
Registration No. 42,079

Date: April 16, 2007

By: Steve Cha
Attorney for Applicant
Registration No. 44,069

IX. CLAIMS APPENDIX

Claim 1. An audio-visual system for processing video data comprising:

an object detection module capable of providing a plurality of object features from the video data, said object features selected from the group of: temporal and spatial feature domains;

an audio processor module capable of providing a plurality of audio features from the video data, said audio features selected from the group consisting of: two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components;

a processor coupled to the object detection and the audio segmentation modules, arranged to determine a maximum correlation value among a plurality of correlation values, wherein each of said correlation values is determined as the sum of the correlation of the selected elements of said audio features and a selected one of the object features.

Claim 2. The system of claim 1, wherein the processor is further arranged to determine whether an animated object in the video data is associated with audio.

Claim 3. (Cancelled)

Claim 4. The system of claim 2, wherein the animated object is a face and the processor is arranged to determine whether the face is speaking.

Claim 5.   The system of claim 4, wherein the plurality of object features are eigenfaces that represent global features of the face.

Claim 6.   The system of claim 1, further comprising:

a latent semantic indexing module coupled to the processor and that preprocesses the plurality of object features and the plurality of audio features before the correlation is performed.

Claim 7.   The system of claim 6, wherein the latent semantic indexing module includes a singular value decomposition module.

Claim 8.   A method for identifying a speaking person within video data, the method comprising the steps of:

receiving video data including image and audio information;

determining a plurality of face image features from one or more faces in the video data, said image features selected from the group of: temporal and spatial feature domains;

determining a plurality of audio features related to audio information, said audio features selected from the group consisting of: two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components;

calculating correlation values, wherein each of said correlation values is determined as the sum of the correlation values of the selected elements of, and each of

said selected face image features; and

determining the speaking person based on a maximum of the correlation values.

Claim 9. The method according to claim 8, further comprising the step of:

normalizing the face image features and the audio features.

Claim 10. The method according to claim 9, further comprising the step of:

performing a singular value decomposition on the normalized face image features and the audio features.

Claim 11. The method according to claim 8, wherein the determining step includes determining the speaking person based upon the one or more faces that has the largest correlation.

Claim 12. The method according to claim 10, wherein the calculating step includes forming a matrix of the face image features and the audio features.

Claim 13. The method according to claim 12, further comprising the step of:

performing an optimal approximate fit using smaller matrices as compared to full rank matrices formed by the face image features and the audio features.

Claim 14. The method according to claim 13, wherein the rank of the smaller

13

matrices is chosen to remove noise and unrelated information from the full rank matrices.

Claim 15.   A memory medium including code for processing a video including images and audio, the code comprising:

code to obtain a plurality of object features from the video, said object features selected from the group of: temporal and spatial feature domains;

code to obtain a plurality of audio features from the video, said audio features selected from the group consisting of: two or more of the following: average energy, pitch, zero crossing, bandwidth, band central, roll off, low ratio, spectral flux and 12 MFCC components;

code to determine correlation values between the plurality of object features and the plurality of audio features, wherein each of said correlation values is determined as the sum of the correlation values of the selected elements of, and each of said selected object features; and

code to determine an association between one or more objects in the video and the audio based on a maximum of the correlation values.

Claim 16.   The memory medium of claim 15, wherein the one or more objects comprises one or more faces.

Claim 17.   The memory medium of claim 16, further comprising code to determine a speaking face.

Claim 18.   The memory medium of claim 15, further comprising:

code to create a matrix using the plurality of object features and the audio features

and   code   to   perform   a   singular   value   decomposition   on   the   matrix.

Claim 19.   The memory medium of claim 18, further comprising:

code to perform an optimal approximate fit using smaller matrices as compared

to   full   rank   matrices   formed   by   the   object   features   and   the   audio   features.

Claim 20.   The memory medium according to claim 19, wherein the rank of the

smaller matrices is chosen to remove noise and unrelated information from the full rank

matrices.

X.   <u>EVIDENCE APPENDIX</u>

No further evidence is submitted herein.

XI.   <u>RELATED PROCEEDING APPENDIX</u>

No related proceedings are pending and, hence, no information regarding same is

available